HHS Public Access

Author manuscript

Comput Methods Programs Biomed. Author manuscript; available in PMC 2020 May 01.

Published in final edited form as:

Comput Methods Programs Biomed. 2019 May ; 173: 167–176. doi:10.1016/j.cmpb.2019.03.002.

Semiparametric Competing Risks Regression Under Interval Censoring Using the R Package intccr

Jun Park,
Department of Biostatistics, Richard M. Fairbanks School of Public Health, Indiana University School of Medicine, 410 W. 10th Street Suite 3000, Indianapolis, IN 46202, United States of America

Giorgos Bakoyannis*,
Department of Biostatistics, Richard M. Fairbanks School of Public Health, Indiana University School of Medicine, 410 W. 10th Street Suite 3000, Indianapolis, IN 46202, United States of America

Constantin T. Yiannoutsos
Department of Biostatistics, Richard M. Fairbanks School of Public Health, Indiana University School of Medicine, 410 W. 10th Street Suite 3000, Indianapolis, IN 46202, United States of America

Abstract

Background and Objective: Competing risk data are frequently interval-censored in real-world applications, that is, the exact event time is not precisely observed but is only known to lie between two time points such as clinic visits. This type of data requires special handling because the actual event times are unknown. To deal with this problem we have developed an easy-to-use open-source statistical software.

Methods: An approach to perform semiparametric regression analysis of the cumulative incidence function with interval-censored competing risks data is the sieve maximum likelihood method based on B-splines. An important feature of this approach is that it does not impose restrictive parametric assumptions. Also, this methodology provides semiparametrically efficient estimates. Implementation of this methodology can be easily performed using our new R package **intccr**.

Results: The R package **intccr** performs semiparametric regression analysis of the cumulative incidence function based on interval-censored competing risks data. It supports a large class of models including the proportional odds and the Fine–Gray proportional subdistribution hazards model as special cases. It also provides the estimated cumulative incidence functions for a particular combination of covariate values. The package also provides some data management

*Correspondence to gbakogia@iu.edu, Phone: +1 317-278-5457.

functionality to handle data sets which are in a long format involving multiple lines of data per subject.

Conclusions: The R package **intccr** provides a convenient and flexible software for the analysis of the cumulative incidence function based on interval-censored competing risks data.

Keywords

interval censoring; competing risks; proportional hazards model; proportional odds model; semiparametric regression; survival analysis

1 Introduction

Competing risk data are time-to-event data where there are multiple mutually exclusive events or causes of failure. The term “competing risks” also includes situations where the scientific interest is focused on class of generalized odds transformation modelsthe first occurring event [1, 2]. In our motivating example, taken from a Human Immunodeficiency Virus (HIV) care and treatment program in sub-Saharan Africa, patients were at risk of death while receiving antiretroviral treatment (ART) and while in care or of becoming lost to care. This latter situation is important because patients who are not retained in care are less likely to receive ART, can infect others in the community and have worse prognosis themselves. In such studies, the interest typically lies on the first event that patients experience, whether this is death or loss to HIV care. The main estimands from such competing risks data are the cause–specific hazard function and the cumulative incidence function. The cause–specific hazard function represents the instantaneous failure rate from a specific event in the presence of the other events, while the cumulative incidence function represents the cumulative probability of an event in the presence of the others. In this article we focus on the analysis of the cumulative incidence function which is the key quantity for studying the risk of occurrence of various events. The cumulative incidence function is used for studying disease prognosis, for evaluating interventions in populations and for prediction and implementation science purposes [3, 4]. In the case of right-censored competing risk data, the packages **cmprsk** and **prodlim** can be used to estimate the cumulative incidence function non-parametrically, based on the Aalen-Johansen estimator [5]. The function cif in the package **compeir** estimates the cumulative incidence function parametrically for each competing risk. For regression analysis of the cumulative incidence function, the packages **cmprsk**, **kmi**, **survival** with the function survfit, and the package **riskRegression** can be used to fit the Fine–Gray proportional subdistribution hazards model [6]. The package **timereg** provides semiparametric estimators for a whole class of models that includes the Fine–Gray model as a special case [7, 8]. Additionally, the package **cmprskQR** performs quantile regression analysis of subdistribution functions [9].

A frequent problem in many clinical studies is that the event time is not precisely observed but is only known to lie between two examination times, such as clinic visits [4, 10–12]. This phenomenon is known as interval censoring in survival and competing risks analysis. In our motivating example, the working definition of loss to care was three months without a clinic visit. This cutoff was chosen by the clinical investigators because, typically, HIV patients receive ART supplies for up to three months at each clinic visit. The analytical

problem is that the exact time of disengagement from HIV care, among patients who have not returned for their next visit, is only known to lie within the three-month interval following the last clinic visit. Similarly, the exact time to death is not known as the data set contains only the death reporting date which is usually after the actual death date. Therefore, the actual death date lies between the last clinic visit of the patient and the death reporting date.

Although interval-censored competing risk data arise frequently in a variety of clinical and medical research settings, only two R packages exist for the analysis of such data. The first is the package **MLEcens**, which applies the height mapping algorithm and the support reduction algorithm by Maathuis [13] and Groeneboom et al. [14] to compute the nonparametric maximum likelihood estimate (NPMLE) of the cumulative incidence function with bivariate interval–censored data. The second is the package **MIICD** which includes the function MIICD.crreg. This package implements the multiple imputation approach proposed by Pan [15] to estimate the regression coefficients and the baseline cumulative incidence function based on the Fine–Gray proportional subdistribution hazards model [6]. However, the package MLEcens does not involve covariates, and the package MIICD uses Rubin's variance estimator, which is well known to be biased when the imputation model and the analysis models are uncongenial [16]. Moreover, the latter package only fits the Fine–Gray proportional subdistribution hazards model [6], and the corresponding regression coefficient estimators are not semiparametrically efficient [17].

The package **intccr** attempts to deal with the aforementioned issues by implementing the semiparametric regression methodology proposed by Bakoyannis, Yu, and Yiannoutsos [4] for the analysis of interval-censored competing risk data. It is important to note that the methodology provides semiparametric efficient regression coefficient estimates [4]. The function ciregic contained in the **intccr** package fits semiparametric regression models for the cumulative incidence function that belong to the large class of generalized odds rate transformation models [18–21] with interval–censored competing risk data. This class includes the Fine–Gray proportional subdistribution hazards model and the proportional odds model as special cases [18]. The function ciregic produces a simple and familiar table of the summarized results. Also, the package **intccr** provides an option for parallel computing that can achieve a substantially faster bootstrap estimation of the variance-covariance matrix for the estimated regression coefficients.

In section 2, the methodological background about the underlying methodology for interval-censored competing risks data is briefly described. Section 3 describes the basic use of the package **intccr** and, also, presents its evaluation through simulation experiments. In section 4, a comprehensive analysis of a real-life data set obtained from an HIV cohort study in sub-Saharan Africa is presented. Future plans and updates are discussed in section 5.

2 Methodology

2.1 Notation

Let T be the actual unobserved event time and $C \in \{1, 2, \cdots, J\}$ be the observed event type or cause of failure. Currently, the package **intccr** allows for two event types, i.e. $C \in \{1, 2\}$.

Let $[a, b]$ denote the observation time interval with $0 < a < b < \infty$. For $i = 1, \cdots, n$, the m_i distinct observation times of the ith study participant are denoted by $a \leq W_{i,1} < W_{i,2} < \cdots < W_{i,mi} \leq b$. Also, the last observation time prior to the event is denoted as V_i and the first observation time after the event as U_i. Based on this notation, the event time of the ith study participant is contained in $(V_i, U_i]$. If the ith study participant's event time is left-censored then $(V_i, U_i] = (0, W_{i,1}]$, if it is right–censored then $(V_i, U_i] = \left(W_{i,m_i}, \infty\right]$, and if it is interval-censored between the observation times $W_{i,k}$ and $W_{i,k+1}$, then $(V_i, U_i] = (W_{i,k}, W_{i,k+1}]$. Now, let $\delta_{ij} = I(V_i < T_i \leq U_i, C = j)$ for $j = 1, 2$ be the indicator function that the ith study participant has experienced the jth event, and the corresponding event time is interval-censored. Similarly, let $\delta_{ij}^1 = I\left(0 < T_i \leq W_{i,1}, C = j\right)$ denote that the ith study participant has experienced the jth event, and the corresponding event time is left-censored. The failure from any event indicator is defined as $\delta_i = \Sigma_{j=1}^2 \left(\delta_{ij} + \delta_{ij}^1\right)$. Obviously, $\delta_i = 0$ indicates that the ith study participant is right–censored. Finally, let $Z \in \mathbb{R}^d$ be a vector of covariates of interest. The observed data for the ith study participant are thus $\boldsymbol{D}_i = \left(V_i, U_i, C_i, \delta_{ij}, \delta_{ij}^1, \mathbf{Z}_i\right)$. The cause–specific cumulative incidence function for the jth event is expressed by

$$F_j(t; \mathbf{z}) = P(T \leq t, C = j \mid \mathbf{Z} = \mathbf{z})$$

for $j = 1, 2$.

2.2 Estimation methodology

With the assumptions that $(W_1, W_2, \cdots, W_m) \perp (T, C)$ conditional on $\mathbf{Z}$ and that the observation time distribution does not contain the parameters of interest (non-informative interval censoring), the likelihood function is

$$L(\boldsymbol{\theta}; \boldsymbol{D}) \propto \prod_{i=1}^{n} \left(\left\{ \prod_{j=1}^{2} \left[F_j\left(U_i; \mathbf{Z}_i, \boldsymbol{\theta}_j\right) - F_j\left(V_i; \mathbf{Z}_i, \boldsymbol{\theta}_j\right) \right]^{\delta_{ij}} \right\} \left\{ \prod_{j=1}^{2} \left[F_j\left(U_i; \mathbf{Z}_i, \boldsymbol{\theta}_j\right) \right]^{\delta_{ij}^1} \right\} \right. \quad (1)$$
$$\left. \times \left[1 - \sum_{j=1}^{2} F_j\left(V_i; \mathbf{Z}_i, \boldsymbol{\theta}_j\right) \right]^{1-\delta_i} \right)$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right)$ are the unknown parameters to be estimated. The cumulative incidence functions can be modeled by using a member of the class of semiparametric transformation models [6, 17, 18] for the cumulative incidence function, which have the general form

$$g_j[F_j(t; z)] = \phi_j(t) + \boldsymbol{\beta}_j^T z$$

for $j = 1, 2$, where $g_j(\cdot)$ is a known increasing link function and $\phi_j(\cdot)$ is an unspecified increasing and invertible smooth function (infinite-dimensional parameter) which is related

to the jth baseline cumulative incidence function. In this case, $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \boldsymbol{\phi}_j)$. A special subset of the class of semiparametric transformation models is the class of generalized odds transformation models which is defined as

$$g_j\left(F_j;\alpha_j\right) = \begin{cases} \log\left[-\log\left(1-F_j\right)\right] & \text{if } \alpha_j = 0 \\ \log\left[\dfrac{\left(1-F_j\right)^{-\alpha_j}-1}{\alpha_j}\right] & \text{if } \alpha_j \in (0,\infty) \end{cases}$$

The Fine–Gray proportional subdistribution hazards model [6] is a special case of this class of models with $\alpha_j = 0$, and so is the proportional odds model [22] with $\alpha_j = 1$. An effective approach to deal with maximum likelihood estimation problems that involve infinite-dimensional parameters is the sieve maximum likelihood approach [23]. This approach avoids some theoretical problems related to likelihood maximization over infinite-dimensional parameter spaces and, also, provides computational efficiency gains [12, 23]. Bakoyannis et al. [4] used a sieve maximum likelihood estimation approach based on B-splines. The corresponding sieve parameter space is given by

$$\mathscr{M}_n\left(\gamma_j, N_j, m_j\right) = \left\{\phi : \phi\left(t;\gamma_j\right) = \sum_{s=1}^{N_j+m_j} \gamma_{j,s} B_{s,m_j}(t), \gamma \in \mathbb{R}^{N_j+m_j}, \gamma_{j,1} < \cdots < \gamma_{j,N_j+m_j}\right\} \quad (2)$$

where N_j and m_j are the number of internal knots and the order of the B-spline for the jth event type or cause of failure, and $\left\{\gamma_j, 1, \cdots, \gamma_{j,N_j+m_j}\right\}$ is the set of B-spline coefficients. For more detailes about the optimal choice of the number of knots see the Discussion Section of this manuscript and Section 2.1 in [4]. Maximizing the likelihood function in Equation (1) with respect to the regression coefficients over a regular Euclidean space and the unspecified functions $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ over the B-spline sieve space provides the sieve maximum likelihood estimates $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\phi}_1, \hat{\phi}_2)$. The consistency for $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\phi}_1, \hat{\phi}_2)$, and the asymptotic normality and semiparametric efficiency of $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$, have been established by Bakoyannis et al. [4].

The function ciregic in the pacakge **intccr** performs the proposed method with nonlinear inequality constraints, using the package **alabama**, to impose the monotonicity constraints involved in Equation (2), which follow from the natural monotonicity of the cumulative incidence function. Additionally, the function ciregic utilizes the package **alabama** to impose the non-linear inequality constraint

$$\max_{\mathrm{z}}\left\{\sum_{j=1}^{2} F_j\left(b;\mathbf{z},\boldsymbol{\theta}_j\right)\right\} < 1,$$

since the sum of the two cumulative incidence functions is a probability and, as such, it is naturally bounded by 1.

3 Basic use of the package and simulation study

The version information of R [24] and the platform of operating system(OS) used in this article are as follows:

```
R> c(R.version$platform, R.version$version.string)
[1] “x86_64-w64-mingw32”   “R version 3.5.2 (2018–12-20)”
```

Under 64-bit version of Windows 10 OS, Monte Carlo simulation and data analysis were performed. With the assumption that the user has the most recent version of R installed, the most recent version of the package **intccr** has to be installed on the user’s OS and loaded as follows:

```
R> install.package(“intccr”)
R> library(intccr)
R> packageVersion(“intccr”)
[1] ‘1.1.1’
```

The package **intccr** provides two simulated data sets. The first data set is longdata which is a long data format, and the second data set is simdata which is in a ready-to-use data format. The data set longdata consists of 200 individuals with 5 variables, where id represents individuals’ identification number, t represents the clinic visit or event evaluation times, c represents the event or censoring indicator, and z1 and z2 are binary and continuous covariates respectively. Note that c has to be 0, 1, or 2, with 0 indicating that the event was not observed throughout the total follow-up period (right censoring). The first 10 observations of longdata are listed below.

```
R> head(longdata, n = 10)
 id     t c z1     z2
1 1 0.86224187 0 0 −2.29032656
2 1 1.20644148 0 0 −2.29032656
3 1 1.73209303 0 0 −2.29032656
4 1 1.73539999 0 0 −2.29032656
5 1 1.96647129 0 0 −2.29032656
6 1 2.12675792 0 0 −2.29032656
7 1 2.46613799 2 0 −2.29032656
8 2 0.05551998 0 1 0.00261902
9 2 0.17492399 0 1 0.00261902
10 2 0.18091429 0 1 0.00261902
```

To analyze the data set longdata in the function ciregic, the data must be reshaped to a suitable format. The package **intccr** provides the function dataprep to reshape data from a long format to a suitable format that is required by the function ciregic.

```
R> newdata <- dataprep(data = longdata, ID = “id”, time = “t”, event = “c”,
Z = c(“z1”, “z2”))
```

The first 10 observations of newdata are given by

```
R> head (newdata, n = 10)
  id    v    u c  z1    z2
1  1 2.1267579 2.4661380 2 0 −2.29032656
2  2 0.1809143 0.3769367 1 1 0.00261902
3  3 2.9436552    Inf 0 1 −1.68379376
4  4 2.4305333    Inf 0 1 −0.90535264
5  5 0.5731781 1.2847889 2 0 0.22854677
6  6 0.0000000 0.3777047 1 0 −0.51449544
7  7 0.0000000 1.4617243 1 1 −1.42043786
8  8 0.0000000 0.4781881 2 1 −0.47006673
9  9 0.1068374 0.9656031 2 0 −0.19349437
10 10 0.3917861 1.0805153 1 0 −0.81510083
R> table (newdata$c)
 0 1  2
29 76 95
```

There are two competing events: the first (c = 1) and the second (c = 2) event type. Right-censored observations are indicated by c = 0. There are 76 observations with the first event type, 95 observations with the second, and 29 observations are right-censored. To elucidate the underlying mechanisms of the function dataprep, Figure 1 shows how longdata is reshaped into newdata via the use of the function dataprep. In longdata, three individuals with id = 1, id = 2, and id = 5 had 7, 4, and 3 time records respectively. These individuals experienced one of the event types between their last two time records. This infers that the event times of those individuals were interval-censored. The function dataprep detected the type of event that an individual experienced and the corresponding time interval. In addition, the function dataprep returned v as the last observation prior to the event and u as the first observation after the event in newdata. The individual with id = 3 who have 8 time records in longdata did not experience any events. The function dataprep returned v = 2.9426552, which is the last time record of the individual with id = 3, as the last observation prior to the event and u = Inf as the first observation after the event in newdata because the individual with id = 3 was right–censored. For the individual with id = 6, the only one time record was observed with event type 1. Therefore, the last observation prior to the event was v = 0 and the first observation after the event was u = 0.3777047 in the newdata because the individual with id = 6 was left-censored. Descriptive statistics for the covariates z1 and z2 in newdata are listed below.

```
R> table (newdata$z1)
```

```
0  1
122 78
R> summary (newdata$z2)
  Min.  1st Qu. Median  Mean 3rd Qu.  Max.
−2.64245 −0.61216 0.02428 0.02383 0.70391 2.86069
```

The arguments of the core function ciregic are described in Table 1. The data must contain the last observation time prior to the event, the first observation time after the event, and the event indicator. The function ciregic fits cumulative incidence models in the class of generalized transformation models on interval–censored competing risk data based on B-spline sieve maximum likelihood estimation. The value of $\boldsymbol{\alpha} = (1, 1)$ for the link functions of the two competing risks is used in this simulation, which corresponds to the proportional odds model [19, 22] for both event types as described in Section 2. This is because the data were simulated from proportional odds models for both event types. Sample R code and the corresponding output of the function ciregic are listed below:

```
R> set.seed (12345)
R> fit.newdata <- ciregic(formula = Surv2(v, u, c) ~ z1 + z2,
           data = newdata, alpha = c(1, 1),
           nboot = 0, do.par = FALSE)
R> fit.newdata
```

Call:

```
ciregic.default(formula = Surv2(v, u, c) ~ z1 + z2, data = newdata,
  alpha = c(1, 1), do.par = FALSE, nboot = 0)
```

Event type 1

Coefficients:

```
   z1    z2
0.5230574 −0.2426299
```

Event type 2

Coefficients:

```
    z1   z2
−0.3963446 0.3442936
```

There are 6 arguments in the function ciregic (see Table 1). The argument formula has the form of response ~ predictor. The response part of the formula must be a Surv2 object in the

function ciregic, and the predictor is a vector of covariates. The first argument in Surv2 is the last examination time before the event, the second is the first examination time after the event, and the last is the event type or censoring status (c ∈ {0, 1, 2}), with 0 indicating right censoring. The argument alpha is a vector of two parameters that represent the link functions of generalized odds rate transformation models for competing events. The support of $\boldsymbol{\alpha}$ is [0, ∞) × [0, ∞). For example, $\alpha_1 = 0$ fits the Fine–Gray proportional subdistribution hazards model [6] for event type 1 and $\alpha_2 = 1$ fits the proportional odds model [22] for event type 2. The argument k is a parameter that controls the number of internal knots of the B-spline. k = 1 is the default, but the user can choose any value satisfying 0.5 ≤ k ≤ 1. Using the half number of internal knots compared to the default can be achieved by choosing k = 0.5 than as the default. This choice can have a substantial effect on computation time with larger data sets. The function ciregic uses cubic B-splines. The argument nboot = 0 forces the function ciregic to returning only the estimated regression coefficients without calculating the bootstrap variance-covariance matrix for the estimated regression coefficients. The function ciregic provides bootstrap variance-covariance matrix for the estimated regression coefficients when a value of the argument nboot is greater than or equal to 2. By setting nboot = 0 and do.par = FALSE, the function ciregic returns only the estimated regression coefficients. This is useful when it is desirable to fit the model and just get point estimates. Below is a sample R code to obtain a bootstrap variance-covariance matrix utilizing parallel computing:

```
R> set.seed(12345)
R> fit.newdata.boot <- ciregic(formula = Surv2(v, u, c) ~ z1 + z2,
             data = newdata, alpha = c(1, 1),
             nboot = 50, do.par = TRUE)
R> summary(fit.newdata.boot)
```

call:

```
ciregic.default(formula = Surv2(v, u, c) ~ z1 + z2, data = newdata,
  alpha = c(1, 1), do.par = TRUE, nboot = 50)
```

Event type 1

```
  Estimate Std. Error z value Pr(>|z|)
z1  0.5231  0.2708  1.931  0.0534 .
z2 −0.2426  0.1067  −2.274  0.0230 *
---
Signif. codes: 0 ‘ ***’ 0.001 ‘ **’ 0.01 ‘ *’ 0.05 ‘ .’ 0.1 ‘ ’ 1
```

Event type 2

```
  Estimate Std. Error z value Pr(>|z|)
z1  −0.3963  0.2509 −1.580  0.11419
z2  0.3443  0.1296  2.657  0.00788 **
---
Signif. codes: 0 ‘ ***’ 0.001 ‘ **’ 0.01 ‘ *’ 0.05 ‘ .’ 0.1 ‘ ‘ 1
```

The argument nboot requires a non-negative integer and denotes the number of bootstrap samples used to estimate a variance-covariance matrix of the estimated regression coefficients. In the above application, we set nboot = 50 and do.par = TRUE. This means that 50 bootstrap samples were used to compute the variance-covariance matrix in parallel computing. The packages **doParallel** and **parallel** are implemented to set the environment for parallel computing, and the package **foreach** is used to perform bootstrap calculations simultaneously. The argument do.par = TRUE detects the number of cores automatically and assigns jobs to the maximum number of available cores. The total number of assigned cores is usually the same as the total number of detected cores minus one.

Extensive Monte Carlo simulations based on 1,000 replications were performed with sample sizes 100, 200, 400, and 800. The results of the simulations are shown in Table 2. The vector of the estimated regression coefficients is $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_{11}, \widehat{\beta}_{12}, \widehat{\beta}_{21}, \widehat{\beta}_{22}\right)'$ which are associated with the estimated regression coefficients of z1 and z2 for the two event types, respectively. Among 1,000 replications for the Monte Carlo simulations, one data set with 100 observations did not converge in at least one bootstrap sample generated in order to calculate the bootstrap standard error. Similarly, two data sets with 200 observations did not converge. Despite these very rare non-convergence issues, the simulation results show negligible bias, similar values of Monte Carlo standard deviation (MCSD) and average standard error (ASE), and values of empirical coverage probability (ECP) close to the nominal level of 0.95. Moreover, the MCSD for the different sample sizes is compatible with a $\sqrt{n}$ convergence rate of the estimator. Figure 2 depicts the true baseline cumulative incidence functions along with the estimated baseline cumulative incidence function for both event types. This Figure illustrates that the function ciregic provides virtually unbiased estimates even with small sample sizes. Table 3 shows summary statistics about the computation times for a single data set in the simulation study. In each scenario, the median computation time using the parallel computing option (do.par = TRUE) to calculate the bootstrap variance-covariance matrix based on 50 bootstrap samples is roughly three times more computationally e cient compared to those without the parallel computing option (do.par = FALSE).

4 Example: Analysis of HIV data using the intccr package

A data analysis from an HIV study on death and loss to HIV care in sub-Saharan Africa is presented in this section. The data were collected by the IeDEA-EA (East African International epidemiology Databases to Evaluate AIDS) Consortium that includes HIV care and treatment programs in Kenya, Uganda, and Tanzania. The data we use here include 3,053 patients who initiated antiretroviral treatment (ART) with a CD4 cell count of at least

100 cells/μl. The data consist of 6 variables, with v being the last clinical examination time prior to the event since ART initiation, u the first clinical examination time after the event, c the event or right-censoring indicator, and age, male and cd4 being the age at ART initiation, male gender indicator, and CD4 cell count at ART initiation, respectively.

```
R> library(intccr)
R> head(iedea, n = 5)
     v     u c   age male cd4
1 0.27104723    Inf 0 35.67146  0 192
2 0.31759068    Inf 0 45.65366  1 191
3 0.14784394 0.1724846 2 62.52977  1 102
4 0.05475701 0.3011636 1 30.77892  0 144
5 2.44490080    Inf 0 43.16496  0 664
```

In total, there were 2,232 patients in HIV care who did not experience any of the events throughout the follow-up period (c = 0, right-censored observations). Moreover, 690 patients were lost to care (c = 1), and 131 patients died while in HIV care (c = 2).

```
R> table(iedea$c)
  0  1  2
2232 690 131
```

Summary statistics regarding age by event type or censoring c are given below:

```
R> tbl.age <- rbind(summary(iedea[iedea$c == 1,]$age, digits = 4),
         summary(iedea[iedea$c == 2,]$age, digits = 4),
         summary(iedea[iedea$c == 0,]$age, digits = 4))
R> rownames(tbl.age) <- c("Loss to care", "Death", "In HIV care")
R> tbl.age
      Min. 1st Qu.  Median Mean 3rd Qu. Max.
Loss to care 18.45  28.63  35.21 36.32  41.65 78.65
Death    20.51  35.16  40.86 42.43  50.75 76.96
In HIV care 18.18  30.42  37.14 38.33  44.87 84.22
```

The median age was 35.2 years, 40.9 years, and 37.1 years for those lost to care, deceased, and still alive and in HIV care at the end of the follow-up period, respectively. Similarly, summary statistics for cd4 by event type are given below:

```
R> tbl.cd4 <- rbind(summary(iedea[iedea$c == 1,]$cd4),
          summary(iedea[iedea$c == 2,]$cd4),
          summary(iedea[iedea$c == 0,]$cd4))
R> rownames(tbl.cd4) <- c("Loss to care", "Death", "In HIV care")
```

```
R> tbl.cd4
     Min. 1st Qu. Median  Mean 3rd Qu. Max.
Loss to care 101 140.25  188 231.8101  262.0 1576
Death     102 131.00  163 187.0687  212.5 1135
In HIV care 101 152.00  199 234.9453  276.0 1332
```

The median CD4 cell count at ART initiation was 188 cells/μl, 163 cells/μl, and 199 cells/μl for those lost to care, deceased, and still alive and in HIV care at the end of the follow-up period, respectively. For the data, we set $\boldsymbol{\alpha} = (1, 1)$, that is we chose the proportional odds model [22] for both event types (i.e. loss to care and death). This choice was made due to the straightforward interpretation of the regression coefficient estimates under the model. For reproducibility purposes regarding the bootstrap variance-covariance matrix of the estimated regression coefficients, we set the seed number to 12345.

```
R> set.seed(12345)
R> fit <- ciregic(formula = Surv2(v, u, c) ~ male + age + cd4,
       data = iedea, alpha = c(1, 1), k = 1, nboot = 50,
       do.par = TRUE)
```

Note that the function factor in the model formula can be used for categorical covariates with more than 2 levels. For example, consider the categorical version of cd4:

$$\text{cd4cat} = \begin{cases} 1 & \text{if cd4} \leq 250 \\ 2 & \text{if } 250 < \text{cd4} \leq 350 \\ 3 & \text{if cd4} > 350 \end{cases}$$

In this case, the analysis can be performed as follows

```
R> set.seed(12345)
R> ciregic(formula = Surv2(v, u, c) ~ male + age + factor(cd4cat),
         data = iedea, alpha = c(1, 1), k = 1, nboot = 50,
         do.par = TRUE)
```

For simplicity, we will use the continuous version of cd4 in the remainder of this Section. The function ciregic is an S3 class function, and therefore the function can be used in conjunction with the generic accessor functions coef, vcov, and summary, as it is illustrated below.

R> summary(fit)

Call:

```
ciregic.default(formula = Surv2(v, u, c) ~ male + age + cd4,
  data = iedea, alpha = c(1, 1), k = 1, do.par = TRUE, nboot = 50)
```

Event type 1

```
   Estimate Std. Error z value Pr(>|z|)
male  0.2128  0.1055  2.017  0.0437 *
age  −0.0295  0.0058  −5.087  <2e−16 ***
cd4  0.0000  0.0003  0.025  0.9797
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Event type 2

```
  Estimate Std.  Error  z  value  Pr(>|z|)
male  0.5668  0.1952  2.904  0.0037 **
age   0.0314  0.0084  3.765  0.0002 ***
cd4  −0.0035  0.0018  −1.989  0.0467 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The computation time for fitting the model and computing standard errors based on 50 bootstraps using the HIV data set of 3,053 individuals was 4.8 minutes with parallel computing and 12.8 minutes without using parallel computing. The function summary returns a summary of the fitted model results with asterisks indicating the corresponding level of statistical significance. The code below extracts the vector of the estimated regression coefficients and the bootstrap variance-covariance matrix, respectively.

```
R> coef(fit)
```

```
R> vcov(fit)
```

The results from the analysis presented above indicate that the odds of loss to care for males is about 24% higher compared to the corresponding odds for females (odds ratio = exp(0.2128) = 1.24). Also, older age at ART initiation by 10 years is associated with a 26% lower odds of loss to care (exp(10 * −0.0295) = 0.74). There is no statistical evidence for an association between CD4 cell count at ART initiation and the cumulative incidence of loss to care as p-value=0.98. Moreover, older age by 10 years is associated with 9% higher odds of death (odds ratio = exp(10 * 0.0084) = 1.09), and, an increased CD4 cell count by 100 cells/μl is associated with 30% lower odds of death (odds ratio = exp(100 * −0.0035) = 0.70). The predicted cumulative incidence functions of loss to care and death for females with a CD4 count of 120 cells/μl at ART initiation, according to age at ART initiation, are depicted in Figure 3. Fitting the proportional subdistribution hazards model (i.e. Fine–Gray model) for

loss to care and the proportional odds model for death can be performed by setting $\alpha_1 = 0$ and $\alpha_2 = 1$, as follows:

```
R> set.seed(12345)
R> fit <- ciregic(formula = Surv2(v, u, c) ~ male + age + cd4,
       data = iedea, alpha = c(0, 1), nboot = 50,
       do.par = TRUE)
```

The generic accessor function predict can be directly used with an object of class ciregic. Table 4 describes the arguments of the function predict. In this example, the argument object is the previously fitted model fit. In the argument covp, the user defines the desired covariate pattern for (male, age, cd4), to predicting the corresponding covariate-specific cumulative incidence functions of loss to HIV care and death. There are 4 lines of output representing 4 different combinations of age by the two event types, “loss to care” (c = 1) and “death” (c = 2) respectively. The argument times produces 100 equally distributed time points between the minimum and the maximum observation time point in the data, for each event type.

```
tms <- fit$tms
par(mfrow = c(1, 2))
t <- seq(from = .2, to = tms[2], by = (tms[2] - .2) / 99)
pred <- lapply(c(20, 30, 40, 50),
      function(x) {predict(object = fit, covp = c(0, x, 120),times = t)})
plot(pred[[1]]$t, pred[[1]]$cif1, type = “l”,
  ylim=c(0, .7), xlim = c(0, 5.5),
  xlab = “Years after ART initiation”,
  ylab = “Cumulative Incidence Function”,
  main = “Loss to care”, lwd = 2)
for(i in 2:4) {lines(pred[[i]]$t, pred[[i]]$cif1, lty = i, col = i, lwd = 2)}
legend(“bottomright”,
   legend = c(“20 years”, “30 years”, “40 years”, “50 years”),
   lty = 1:4, col = 1:4, lwd = rep(2, 4))
t <- seq(from = tms[1], to = tms[2], by = diff(tms) / 99)
pred <- lapply(c(20, 30, 40, 50),
      function(x) {predict(object = fit, covp = c(0, x, 120),times = t)})
plot(pred[[1]]$t, pred[[1]]$cif2, type = “l”,
  ylim=c(0, .1), xlim = c(0, 5.5),
  xlab = “Years after ART initiation”,
  ylab = “Cumulative Incidence Function”,
  main = “Death”, lwd = 2)
for(i in 2:4) {lines(pred[[i]]$t, pred[[i]]$cif2, lty = i, col = i, lwd = 2)}
legend(“bottomright”,
   legend = c(“20 years”, “30 years”, “40 years”, “50 years”),
```

```
   lty = 1:4, col = 1:4, lwd = rep(2, 4))
par(mfrow = c(1, 1))
```

Moreover, the waldtest function can be used to perform a Wald test based on an object from the the function ciregic in the package **intccr**. Below are three examples of performing a Wald test. In the first example we compare a model (male, age, cd4) with the null model (model without covariates).

```
R> set.seed(12345)
R> fit.f <- ciregic(formula = Surv2(v, u, c) ~ male + age + cd4,
         alpha = c(1, 1), nboot = 50, do.par = TRUE,
         data = iedea)
R> waldtest(full = fit.f)
```

Full model: male age cd4

Nested model:

Wald test

```
 Chisq df P(> Chisq)
75.5101 6  3e-14
```

Wald test (cause-specific)

Event type 1

```
 Chisq df P(> Chisq)
26.1007 3  9e-06
```

Event type 2

```
 Chisq df P(> Chisq)
39.5401 3  1e-08
```

The function waldtest returns output for two parts: one is the test for the effect of the covariates on any event type and the other is the event-specific test. In the above example, the χ^2 statistic of overall test is 75.5 and its p-value is close to 0. Also, the χ^2 statistic of each test for event types 1 and 2 is 26.1 and 39.5 respectively and those p-values are close to 0. These results indicate that the variables male, age, and cd4 should be in the model because parameters associated with those variables are not zero. The next example is the Wald test comparing a model with covariates male, age, and cd4 to the nested model with covaraites male and age.

```
R> fit.n <- ciregic(formula = Surv2(v, u, c) ~ male + age,
         alpha = c(1, 1), nboot = 50, do.par = TRUE,
         data = iedea)
R> waldtest(full = fit.s, nested = fit.n)
```

Full model: male age cd4

Nested model: male age

Wald test

```
 Chisq df P(> Chisq)
4.0879 2  0.1295
```

Wald test (cause-specific)

Event type 1

```
 Chisq df P(> Chisq)
6e-04 1  0.9797
```

Event type 2

```
 Chisq df P(> Chisq)
3.9551 1  0.0467
```

The χ^2 statistic for the effect of CD4 of the cumulative incidence of death is 3.96 with p-value = 0.047, indicating a statistically significant effect of CD4 on the cumulative incidence of death.

```
R> fit.n1 <- ciregic(formula = Surv2(v, u, c) ~ age,
         alpha = c(1, 1), nboot = 50, do.par = TRUE,
         data = iedea)
R> waldtest(full = fit.s, nested = fit.n1)
```

Full model: male age cd4

Nested model: age

Wald test

```
 Chisq df P(> Chisq)
23.7261 4  9e-05
```

Wald test (cause-specific)

Event type 1

```
 Chisq df P(> Chisq)
4.1325 2  0.1267
```

Event type 2

```
 Chisq df P(> Chisq)
11.1499 2  0.0038
```

In the above example, the χ^2 statistic of overall test is 23.7 and its p-value is close to 0. Also, the χ^2 statistic of each test for event types 1 and 2 is 4.1 and 11.1 respectively and those p-values are 0.12 and close to 0 respectively.

5 Discussion

The package **intccr** provides a convenient and versatile tool for robust semi-parametric regression analysis of the cumulative incidence function based on interval–censored competing risk data. The package supports a large class of models for the cumulative incidence function, including the proportional odds and the Fine–Gray proportional subdistribution hazards model as special cases. It also provides semiparametrically efficient regression coefficient estimates. To the best of our knowledge, the only other available software for the analysis of interval-censored competing risks data is the R package **MIICD**. That package utilizes Rubin's multiple imputation approach to deal with the unobserved event times. However, it is well known that Rubin's variance estimator is biased in cases where the imputation and the analysis models are uncongenial [16], a scenario that occurs frequently in practice. In addition, the **MIICD** package does not provide semiparametrically efficient regression coefficient estimates and it only supports the Fine–Gray proportional subdistribution hazards model [6], whose interpretation is more difficult compared to the proportional odds model. The package **intccr** follows the guideline for selecting the number of knots in Section 2.1 in Bakoyannis, Yu, and Yiannoutsos [4]. Briefly, the number of internal knots for the B-spline is $N = \lfloor k \times n^{1/3} \rfloor$ where $\lfloor a \rfloor$ is the largest integer that is smaller than or equal to the real number a, $k \in [0.5, 1]$ is a parameter that is specified by the user and n is the sample size. For more details about the justification of the selection of knots, please see Section 2.1 in Bakoyannis, Yu, and Yiannoutsos [4]. The package **intccr** uses flexible cubic B-splines which is a standard choice in practice. Regarding the maximum number of regression coefficients to be estimated (or equivalently the maximum number of covariates) for each event type, we suggest the following rule of thumb:

$$\text{Maximum number of covariates} = \left\lfloor \frac{\min(n_1, n_2)}{10} \right\rfloor$$

where n_j, $j = 1, 2$, is the number of observations with event type j.

It has to be noted that, in many cases, there is no obvious interval censoring. However, the event time is typically measured in days, and the exact time of the event occurence is not recorded. In this case, assuming that the true event time is a continuous, the exact event time is still interval censored, with the width of the censoring interval being 1 day. Such cases, can still be analyzed using the package **intccr** by setting $V = X - 0.5$ days and $U = X + 0.5$ days, where X is the recorded event time in days. This occurs, for example, in Dementia studies, where the time to Dementia is interval-censored while the time to death is more precisely recorded in days. Such data can be easily analyzed using the package **intccr**.

The simulations were run on Intel(R) Core(TM) i5–2400 CPU 3.10GHz with 8 GB ram. Maximum number of available cores in parallel computing was 3. We expect that the users having higher specification of their computer may see more timely efficient results.

The **intccr** package introduced in this paper provides the estimated cumulative incidence functions for a particular combination of covariate values. This quantity is very appealing for graphical illustration. Also, the package provides data management functionality to reformat data sets provided in a long format (i.e. data sets with multiple lines per subject), and turn them into the wide (single-line per subject) format required by the package. One limitation of the **intccr** package is that, for the time being, it only allows for two event types or causes of failure. We plan to update our package to allow for more than two event types in the near future. The package is freely available for download from the CRAN website https://cran.r-project.org/web/packages/intccr/index.html.

Ackowledgments

We thank the Editor-in-Chief, Associate Editor and the three anonymous reviewers for their helpful comments that led to a significant improvement of this manuscripts and the intccr package. This research was supported by the National Institute of Allergy and Infectious Diseases (NIAID), Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD), National Institute on Drug Abuse (NIDA), National Cancer Institute (NCI), and the National Institute of Mental Health (NIMH), in accordance with the regulatory requirements of the National Institutes of Health under Award Number U01AI069911 East Africa IeDEA Consortium. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research has also been supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through USAID under the terms of Cooperative Agreement No. AID-623-A-12–0001 it is made possible through joint support of the United States Agency for International Development (USAID). The contents of this presentation are the sole responsibility of AMPATH and do not necessarily reflect the views of USAID or the United States Government.

References

[1]. Putter H, Fiocco M, and Geskus RB. Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine, 26(11):2389–2430, 2007. doi: 10.1002/sim.2712. [PubMed: 17031868]

[2]. Bakoyannis Giorgos and Touloumi Giota. Practical methods for competing risks data: A review. Statistical Methods in Medical Research, 21(3):257–272, 2012. doi: 10.1177/0962280210394479. [PubMed: 21216803]

[3]. Koller Michael Thodor, Raatz Heike, Steyerberg Ewout Willem, and Wolbers Marcel. Competing risks and the clinical community: irrelevance or ignorance? Statistics in medicine, 31(11–12): 1089–1097, 2012. doi:10.1002/sim.4384. [PubMed: 21953401]

[4]. Bakoyannis Giorgos, Yu Menggang, and Yiannoutsos Constantin T.. Semi-parametric regression on cumulative incidence function with interval-censored competing risks data. Statistics in Medicine, 36(23):3683–3707, 2017. doi: 10.1002/sim.7350. [PubMed: 28608412]

[5]. Aalen Odd O. and Johansen Søren. An empirical transition matrix for non-homogeneous markov chains based on censored observations. Scandinavian Journal of Statistics, 5(3):141–150, 1978 URL http://www.jstor.org/stable/4615704.

[6]. Fine Jason P. and Gray Robert J.. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association, 94(446):496–509, 1999. doi: 10.1080/01621459.1999.10474144.

[7]. Scheike Thomas H., Zhang Mei-Jie, and Gerds Thomas A.. Predicting cumulative incidence probability by direct binomial regression. Biometrika, 95(1):206–220, 2008. doi: 10.1093/biomet/asm096.

[8]. Scheike Thomas H. and Zhang Mei-Jie. Analyzing competing risk data using the r timereg package. Journal of Statistical Software, 38(2), 2011. doi: 10.18637/jss.v038.i02.

[9]. Peng Limin and Fine Jason P.. Competing risks quantile regression. Journal of the American Statistical Association, 104(488):1440–1453, 2009. doi:10.1198/jasa.2009.tm08228.

[10]. Sun Jianguo. The Statistical Analysis of Interval-censored Failure Time Data. Springer-Verlag New York, 2006.

[11]. Chen Ding-Geng, Sun Jianguo, and Peace Karl E.. Interval-censored time-to-event data: methods and applications. Chapman & Hall/CRC Biostatistics Series, 2012.

[12]. Zhang Ying, Hua Lei, and Huang Jian. A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. Scandinavian Journal of Statistics, 37(2):338–354, 2010. doi: 10.1111/j.1467-9469.2009.00680.x.

[13]. Maathuis Marloes H.. Reduction algorithm for the npmle for the distribution function of bivariate interval-censored data. Journal of Computational and Graphical Statistics, 14(2):352–362, 2005. doi: 10.1198/106186005X48470.

[14]. Groeneboom Piet, Jongbloed Geurt, and Wellner Jon A.. The support reduction algorithm for computing non-parametric function estimates in mixture models. Scandinavian Journal of Statistics, 35(3):385–399, 2008. doi: 10.1111/j.1467-9469.2007.00588.x.

[15]. Pan Wei. A multiple imputation approach to cox regression with interval-censored data. Biometrics, 56(1):199–203, 2000. doi: 10.1111/j.0006-341X.2000.00199.x. [PubMed: 10783796]

[16]. Meng Xiao-Li. Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 9(4):538–558, 1994. doi: 10.1214/ss/1177010269.

[17]. Mao Lu and Lin Dan-Yu. E cient estimation of semiparametric transformation models for the cumulative incidence of competing risks. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(2): 573–587, 2017. doi: 10.1111/rssb.12177. [PubMed: 28239261]

[18]. Jeong Jong-Hyeon and Fine Jason P.. Parametric regression on cumulative incidence function. Biostatistics, 8(2):184–196, 2007. doi: 10.1093/biostatistics/kxj040. [PubMed: 16636138]

[19]. Dabrowska Dorota M. and Doksum Kjell A.. Estimation and testing in a two-sample generalized odds-rate model. Journal of the American Statistical Association, 83(403):744–749, 1988 URL http://www.jstor.org/stable/2289300.

[20]. Fine Jason P.. Regression modeling of competing crude failure probabilities. Biostatistics, 2(1): 85–97, 2001. doi: 10.1093/biostatistics/2.1.85. [PubMed: 12933558]

[21]. Scharfstein Daniel O., Tsiatis Anastasios A., and Gilbert Peter B.. Semi-parametric e cient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. Lifetime Data Analysis, 4(4):355–391, 1998. [PubMed: 9880995]

[22]. Eriksson Frank, Li Jianing, Scheike Thomas, and Zhang Mei-Jie. The proportional odds cumulative incidence model for competing risks. Biometrics, 71(3):687–695, 2015. doi: 10.1111/biom.12330. [PubMed: 26013050]

[23]. Shen Xiaotong and Wong Wing Hung. Convergence rate of sieve estimates. The Annals of Statistics, 22(2):580–615, 1994.

[24]. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2018 URL http://www.R-project.org/.

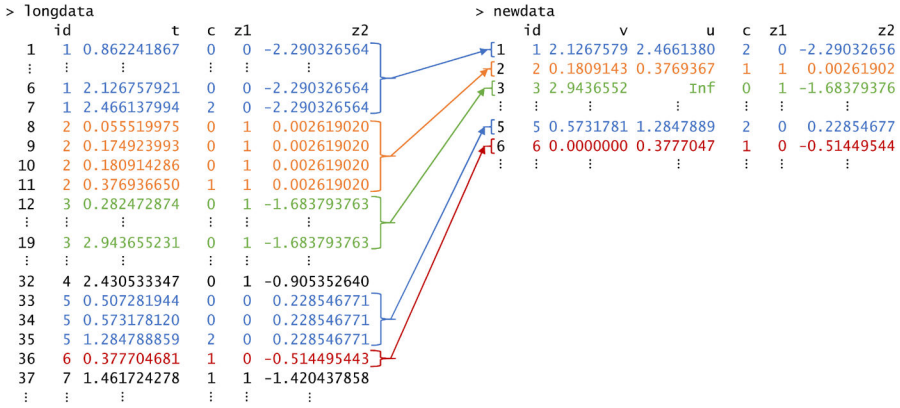
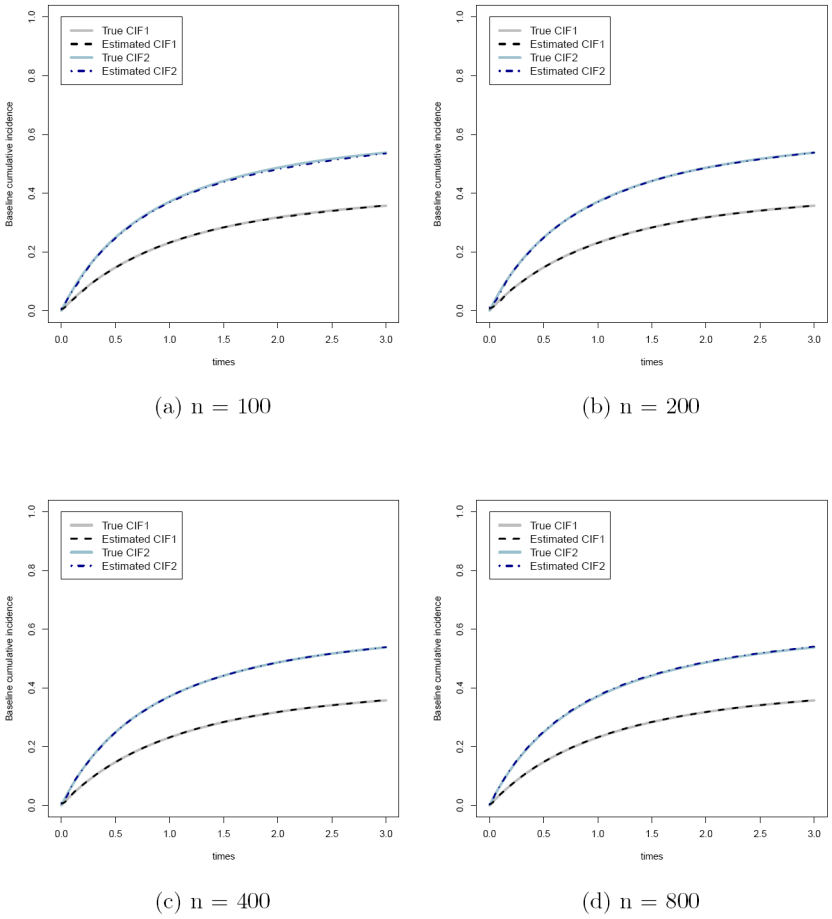


Figure 1:
Data resphaping with the function dataprep



(a) n = 100 (b) n = 200

(c) n = 400 (d) n = 800

Figure 2:

Baseline cumulative incidence function Solid gray and light blue lines indicate true baseline cumulative incidence functions of the event type 1 and the event type 2 respectively. Dotted black and blue lines indicate the estimated baseline cumulative incidence functions of the event type 1 and the event type 2 respectively.

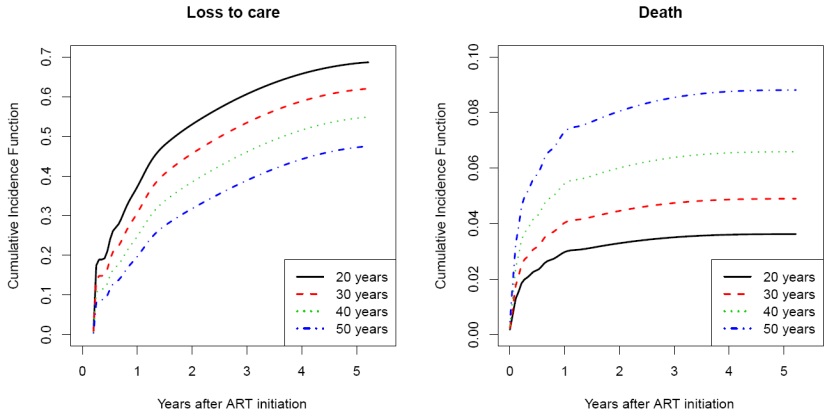


Figure 3:
Predicted cumulative incidence functions for females aged 20 to 50 years, with CD4 count 120 cells/μl at ART initiation

Table 1:

Arguments in the function ciregic

Arguments	Description
formula	a formula object relating survival object Surv2(v, u, event) to a set of covariates
data	an input data frame
alpha	parameters that define the link functions from class of generalized odds-rate transformation models
k	a parameter that controls the number of internal knots in the B-spline with k∈ [0.5, 1]
nboot	a number of bootstrap samples for estimating variances and covariances of an estimated regression coefficients
do.par	a logical constant for using parallel computing for bootstrap variance estimation

Table 2:

Monte Carlo simulation results based on 1,000 replications The standard error is estimated by bootstrap sampling. Monte Carlo standard deviation (MCSD), average standard error (ASE), empirical coverage probability (ECP).

n	Event type	Parameters	%bias	MCSD	ASE	ECP
100	1	β_{11}	−3.055	0.403	0.425	0.957
		β_{12}	4.645	0.189	0.207	0.967
	2	β_{21}	−1.033	0.394	0.417	0.960
		β_{22}	4.867	0.191	0.202	0.961
200	1	β_{11}	−0.578	0.282	0.285	0.954
		β_{12}	2.983	0.144	0.140	0.939
	2	β_{21}	1.418	0.273	0.282	0.948
		β_{22}	2.737	0.139	0.136	0.936
400	1	β_{11}	0.683	0.198	0.197	0.947
		β_{12}	−0.851	0.097	0.097	0.952
	2	β_{21}	1.812	0.196	0.195	0.953
		β_{22}	−0.127	0.095	0.095	0.946
800	1	β_{11}	2.160	0.138	0.138	0.951
		β_{12}	0.261	0.070	0.069	0.941
	2	β_{21}	1.884	0.136	0.136	0.946
		β_{22}	−0.011	0.066	0.067	0.944

Table 3:

Computation times (seconds) for fitting the model and calculating the standard errors using 50 bootstrap samples based on Monte Carlo simulation with 1,000 replications

Parallel	Yes (do.par = TRUE)					No (do.par = FALSE)				
n	**Min**	**Q1**	**Median**	**Q3**	**Max**	**Min**	**Q1**	**Median**	**Q3**	**Max**
100	29.78	37.45	39.96	42.80	57.70	81.92	102.22	108.74	116.00	145.39
200	29.85	38.59	41.23	43.96	65.06	85.59	105.00	112.94	120.80	166.98
400	31.91	42.59	45.95	49.80	70.07	88.70	116.04	126.58	136.70	194.84
800	39.33	49.53	52.95	56.83	75.22	105.36	135.38	145.95	156.90	209.63

Table 4:

The arguments of the function predict

Arguments	Description
object	An object of class ciregic, generated from the fitted model
covp	The vector of covariates
Times	User-defined time points used to predict the cumulative incidencefunctions